DOCUMENT RESUME

ED 420 670 TM 028 345

AUTHOR	Goldberg, Gail Lynn; Roswell, Barbara Sherr
TITLE	Perception and Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment.
PUB DATE	1998-04-00
NOTE	54p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150) -- Tests/Questionnaires (160)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Criteria; Educational Assessment; Educational Practices; Elementary Secondary Education; Experience; *Performance Based Assessment; Professional Development; *Scoring; Standardized Tests; *State Programs; *Teacher Attitudes; Teaching Methods; *Testing Programs; Thinking Skills
IDENTIFIERS	Charles County Public Schools MD; *Maryland School Performance Assessment Program

ABSTRACT

Teachers' reactions to the administration and scoring of the Maryland School Performance Assessment Program tests (MSPAP) were studied, focusing on their direct and indirect exposure to tasks and evaluative criteria through the experience of scoring the MSPAP. Since its inception in 1991, the MSPAP has been scored in-state by certified teachers from Maryland. Many teachers have identified the opportunity to score the MSPAP as an opportunity for professional development and a chance to familiarize themselves with the test and its objectives. About 50 teachers from Charles County (Maryland) completed questionnaires about the impact of scoring the MSPAP on their teaching and their perceptions of how the MSPAP is integrated into their own and their colleagues' instructional practices. Twelve Charles County teachers (experienced scorers) from four different schools were also interviewed about the impact of MSPAP. Almost without exception, teachers endorsed the scoring experience as one that galvanized them and made them more reflective, critical, and deliberate. Thanks largely to their scoring experience, they perceived their own classroom activities as more likely to elicit writing for varied and coherent purposes, to integrate content, and to cue for higher order thinking. However, teachers note that the scoring experience does not provide them with a well-grounded understanding of performance assessment. This finding supports the view that tests alone will not result in improved instruction overall without well-planned staff development. An appendix contains sample interview questions. (Contains 14 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 670

Perception and Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Gail Goldberg
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Gail Lynn Goldberg

Educational Consultant for Charles County Public Schools

Barbara Sherr Roswell

Goucher College

TM028345

Paper presented at the annual meeting of the American Educational Research Association,

April 13-17, 1998, San Diego, California

Perception and Practice: The Impact of Teachers' Scoring Experience on Performance-Based Instruction and Classroom Assessment

Introduction

Increasingly, professional conversation within the educational assessment community about the impact of large-scale, standardized administration of performance assessment tests is being directed towards the consequential aspects of validity. Perspectives range from fears about "teaching to the test" to confidence in the capacity of these assessments to model, support, and positively shape curricular and instructional reform. When, as in the case of the Maryland School Performance Assessment Program (MSPAP) tests, teachers claim that the test has had a favorable impact on instruction (Waldron, 1997), we ought to be moved to investigate these claims and to explore the empirical support for them.

Responding to the injunction in a 1997 AERA forum that in order to examine the consequences of testing we extend the kinds of questions and types of evidence we consider (Moss, 1997), we investigated teachers' reactions to the administration and scoring of MSPAP in a range of contexts. We explored how teachers "take testing home," interpret and respond to the curricular and instructional approaches modeled by the test, and apply these interpretations to instructional practices and materials. Through this study we sought to discover to what extent, if at all, classroom practices are actually changing, and in what ways. What aspects of exposure to, and increasing familiarity with, performance assessments are stimulating change? Moving from Wiggins' argument that the validity of a test cannot be evaluated apart from the kind of instruction it is intended to support (1992), we must also explore whatever discrepancies may exist between the instruction that performance assessment is intended to support and that which,

it actuality, it may be supporting. The purpose of this study was to address these and other related questions by focusing on the impact of one vehicle through which performance assessment has purportedly improved instruction--teachers' direct and indirect exposure to tasks and evaluative criteria through the experience of scoring the MSPAP.

Background

Since its inception in 1991, MSPAP has been scored in-state by certified teachers who reside and/or teach in Maryland. Aside from practical reasons for participating in the five to six week long project (e.g., to supplement earnings and to earn continuing education credit), many teachers identify the experience as one which provides professional development not otherwise available through system- or state-based activities. So highly regarded is the experience that a number of Maryland's twenty-four local educational agencies (LEAs) have attached financial and other incentives for teacher participation and have lobbied vigorously for the opportunity, which rotates periodically, to host one of the four regional sites at which summer scoring occurs each year. Despite the widespread enthusiasm, however, there has been little actual evidence accrued to date of how (and indeed, if) teachers effectively apply training and experience in scoring performance tasks to their own classroom practice (see Afflerbach, Guthrie, Schafer, & Almasi, 1994; Koretz, Mitchell, Barron, & Keith, 1996).[1]

One LEA that actively pursued and succeeded in being selected in 1995 as a scoring site

---

[1] Through a grant from the U. S. Department of Education, The Maryland State Department of Education has funded a consequential validity study conducted by Dr. Suzanne Lane and colleagues at the University of Pittsburgh in collaboration with MSDE. Now in its third year, the preliminary report is due to be released in content-specific sections, beginning later this spring. No data from this examination of the impact of MSPAP on schools and local school systems is available at this time.

is Charles County. A rapidly growing and changing district with a student population of nearly 22,000, this system sought a scoring site specifically for the perceived benefits it would yield in terms of exposure to, and experience with, judgment-based scoring of performance tasks. Since the summer of 1995, increasing numbers of Charles County teachers have availed themselves of the opportunity to score MSPAP. As a follow-up to summer scoring of the '96 MSPAP, in the autumn of 1996 Charles County implemented a system-wide, day-long in-service program on scoring MSPAP, utilizing as trainers county teachers who had scored MSPAP and in many cases had served as scoring coordinators or team leaders. Using several "public release tasks" (actual MSPAP tasks and scoring guides used in past editions of the test), all elementary and middle school teachers and instructional leaders (principals, assistant principals, etc.) were trained on judgment-based scoring through the application of MSPAP rubrics and activity-specific scoring tools. Because of the tight security during operational testing, for many educators this was the first opportunity to see a complete MSPAP task rather than a mere prototype or sample task.

Operational scoring of MSPAP is an activity for which teachers must apply. Those who are selected are each assigned to a team based on grade level and content area expertise and interest. Because of site size and location, teachers in Charles County were assigned to score either Grade 3 or Grade 8. Operational scoring training takes place over a two to three day period, during which all participants must qualify by reaching 70% exact agreement with pre-established "true scores" on one or more qualifying sets of student responses. Teams are each responsible for scoring approximately one-fourth of the items in a given cluster or grade-level edition of the test. Over the course of a summer, team members each score approximately one thousand booklets, thus gaining extensive exposure to a limited number of items, usually

measuring two, and no more than three, different content areas and only some outcomes within those areas.

Data Sources and Methods

The scoring experience provides an opportunity for teachers to see first-hand the relationships among: (1) the Maryland Learning Outcomes (MLOs), which identify what students are supposed to have learned and be able to do in reading, writing, language usage, mathematics, science, and social studies; (2) a sequence of activities comprising one or more complex, often integrated performance tasks; and (3) the evaluative criteria by which evidence of proficiency in the MLOs is judged. To assess the impact of this experience, we therefore identified, obtained, and analyzed several types of evidence of teachers' understanding and application of those relationships which could be linked to exposure (both direct and indirect) to performance assessment through judgment-based scoring.

In late spring, prior to the scoring of the 1997 MSPAP, we obtained samples of instructional activities and classroom assessments used during the 1996-97 school year. Some materials were designed by teachers with first-hand experience scoring MSPAP, and the rest by teachers who had experienced the countywide in-service training on scoring. We conducted a close analysis of these classroom instructional and assessment activities, focusing on 1) their alignment with the learning outcomes which underlie MSPAP design, scoring, and reporting of data, as well as the alignment between county curriculum frameworks (which are also supposed to be embodied in all instructional activities) and the MLOs; 2) the characteristics of activities/tasks in terms of cuing and format; and 3) the clarity and appropriateness of evaluative criteria and assessment strategies.

Shortly after the onset and again at the conclusion of operational scoring in the summer of 1997, we administered to approximately 50 Charles County teacher-scorers a pair of questionnaires on the impact of scoring MSPAP on their teaching and their perceptions of how MSPAP is integrated into their own and their colleagues' instructional and classroom assessment practices. To gain insight into how the scoring experience might change attitudes and understandings, respondents were asked to: 1) define performance-based instruction and assessment; 2) identify personal and school-level needs in terms of information about such topics as the MLOs, task design, and developing scoring criteria; 3) indicate their degree of familiarity with available instructional support resources; 4) rate the impact of the scoring experience (past and current) on their knowledge and practice; and 5) describe things they had done (or intended to do) differently in their classroom as a result of the scoring experience. Questionnaire data were compared to, and considered in light of, the sample materials earlier obtained from these individuals and their colleagues, to illuminate similarities and differences between perceived and actual practice. Additional instructional and classroom assessment activities were obtained and examined in the months that followed, once these teachers returned to their classroom and had the opportunity to implement anticipated changes.

Finally, to expand our understanding of the wider impact of teachers' scoring experience on their own practice and that of their immediate instructional communities, we conducted interviews with twelve Charles County teachers from four different schools who had one or more years' scoring experience (see Appendix A for interview questions) and undertook informal classroom observations. Hypothesizing that teachers did not follow a simple linear path from the scoring experience to instructional delivery, we sought to understand factors that

supported or impeded teachers' attempts to "put MSPAP into practice" and to be mindful of possible differences between teachers' stated perceptions and goals and the ways these ideas might or might not be translated into specific instructional and classroom assessment activities.

Findings

Questionnaire Responses

The questionnaires were designed as conversation opener, and posed a range of questions that provided insight into 1) teacher-scorers' familiarity both with the terminology associated with performance-based instruction and assessment and with resources that might support them in creating more performance-based classrooms; 2) teachers' perceived needs; and 3) the ways teachers understood and planned to respond to the scoring experience. Unlike situations in which questionnaires go out like "cold calls," respondents saw themselves as part of a community that included the researchers and knew that the information and perspectives they expressed were part of an ongoing dialogue. Without exception, their responses were forthcoming and candid.

Definitions of terminology

In order to better understand the degree to which teachers were familiar with performance-based instruction and performance assessment and to chart changes in their understandings as a result of the scoring experience, both at the beginning and at the end of operational scoring, teachers were asked to define these two terms in their own words. Their wide range of responses revealed that while scorers certainly have a general understanding of performance based instruction as a form of teaching in which students learn by doing "real life tasks," their familiarity with these concepts is often partial, hodge-podge, or superficial. A very

small number of responses revealed a seriously flawed definition of one or both of these terms (for example, performance-based instruction "is a non-content related method of teaching" in which "the teacher does not really teach") or a confused linking of terms (performance based-instruction is "instruction based on demonstrating a task. Show how a procedure is followed. Provide strategies that teach students to follow a sequential order of steps"). A more general pattern, however, was for teachers either to conflate instruction and assessment or to highlight certain elements of MSPAP-like tasks at the expense of others, suggesting that performance-based instruction could be defined by one or two of its key elements (use of hands-on activities, integrated content, emphasis on higher order thinking, inclusion of group and peer work, application of knowledge to real-world situations, etc.). Conspicuously absent were references either in the questionnaires, or later in interviews, to prominent research and resources in the field, or to any comprehensive, theoretical rationale for embracing performance based instruction (with only one teacher using the term "discovery learning" to identify the approach on which MSPAP was based). When asked later in interviews explicitly if their teacher training had prepared them in any way for performance-based instruction or classroom assessment, teachers, with the exception of one 1997 graduate, said no.

Teachers' ratings of knowledge about performance-based instruction and assessment

Given the variety of definitions of performance-based instruction and performance assessment in the questionnaires, we were particularly curious to learn how teachers would assess their own and their colleagues' knowledge and use of performance-based instruction and performance assessment before scoring, and also to assess their individual knowledge now that scoring was complete. Teachers therefore were asked to rate their own and others' knowledge as

lacking, limited, moderate, or considerable.[2]

Overall, the 37 teachers who completed both the initial and final questionnaire rated their own knowledge of performance instruction and performance assessment before scoring, like that of their colleagues, to be limited (approximately 2.7 on a 5 point scale), and rated the school administrators' knowledge as only slightly greater (3.0). After scoring, however, they identified their knowledge as moderate or considerable (3.6), now outstripping both colleagues' and administrators' expertise (see Table 1). Teachers similarly said that although before scoring their use of performance-based instructional activities and performance assessment was limited (2.9 and 2.7, respectively), they predicted that following scoring their use of these approaches would be moderate or considerable (3.7 and 3.6). While in most categories there was no significant difference in ratings between the 20 teachers who had scored for one year only and the 17 others who had scored for two or more years, those who had scored for multiple years rated both their own knowledge of performance based instruction as a result of scoring and their expectations for using performance-based activities and performance assessments slightly higher than did those new to scoring (knowledge 3.47 first year, 3.71 multiple years; use of activities 3.60, 3.76; and use of performance assessment 3.55, 3.76).

Expectations for changed instruction as a result of scoring

Questionnaires suggested a cluster of ways that teachers most often reported that they had changed or would change their teaching as a result of the scoring experience. These included (in order of frequency):

---

[2] The four-element Likert-type scale was converted to a five-element scale with a "dummy middle" in order to facilitate analysis with SPSS. Responses were converted to numerical value, with lacking=1, limited=2, moderate=4, and considerable=5.

- incorporating more performance activities in their classrooms and creating hands-on activities aligned with the MLOs
- using more or better rubrics in classroom assessment
- assigning more writing, and specifically, more writing for a variety of purposes and in a variety of content areas
- encouraging students to explain and elaborate their answers and to return to resources for evidence
- putting more emphasis on reading and writing skills, specifically including more non-fiction selections
- including more "MSPAP vocabulary" in everyday teaching and "teaching MSPAP as part of regular curriculum"
- attempting more content integration
- focusing on students' self-assessment, problem analysis and problem solving
- including more timed tasks
- consciously focusing on indicator level concepts and skills (e.g., organizing and displaying data in graphs, symmetry, critical stance in reading)
- placing a higher value on careful work (whether encouraging students to check spelling and punctuation or, as one teacher vowed, "to beat profusely any student who does not put a title on his or her graph")
- changing classroom management to include more group and team work and more movement in and around the classroom

Only one teacher said that she expected to make "no changes" in her teaching as a result of

scoring.

Teachers' responses changed in predictable patterns from the outset to the conclusion of scoring. Initial responses more often focused on "easy fixes," gimmicks, and quick tips for avoiding simple errors and thus improving their students' scores, such as reminding students to label their graphs. At this time, teachers also often noted their plans to incorporate more "MSPAP vocabulary" in their teaching. By the end of scoring, teachers tended to identify more global ways they were reconceptualizing their teaching that were less concerned with the test itself, and more concerned with such issues as content integration, using writing across the curriculum, and helping students to assess themselves more accurately and reflect on and explain their thinking processes. Even from the beginning to the end of scoring, there was recognizable movement from a narrow concern with teaching to the test to using what they had learned from scoring the test to inform and improve teaching. In interviews that followed, their comments revealed that veteran scorers were more likely than neophytes to have made this transition.

Familiarity with MSPAP-related resources and requests for additional information and support

Although teachers attributed important changes in their knowledge and practice to the scoring experience, these changes may be best understood in the wider context of what they already knew about MSPAP and performance-based instruction and assessment. Therefore, in the initial questionnaire, teacher-scorers were asked to indicate their degree of familiarity with six different resources which had been developed and disseminated (at the system or school, but not teacher level) by the Maryland State Department of Education to increase understanding of

MSPAP and the larger instructional objectives the program is intended to support. Specifically, they were asked to indicate if they used, have seen but don't use, know about but have not seen, or have never heard about the following: public release tasks (of which there were fourteen at that time); MSPAP writing and language usage rubrics; Teacher to Teacher Talk (an annual collection of scorers' observations about students' responses to MSPAP and their instructional implications published from 1992-1996); Scoring MSPAP: A Teacher's Guide (an overview which includes sample items and scoring tools for each content area); MSPAP Exemplars (models of performance-based lessons); and MSPAP Clarifications Documents (content area-specific elaborations on the definitions and ways of addressing the MLO indicators; Social Studies had been released only months earlier and both Mathematics and Science had only limited circulation in draft form). Of the six resources, respondents indicated the greatest familiarity with and use of the rubrics (25) and Teacher to Teacher Talk (24). While a large number also indicated use of the public release tasks (20) and Scoring MSPAP: A Teacher's Guide (18), a significant number (9, 7) indicated that they had never even heard of these documents. It was not surprising that almost no teachers (6) were even aware of the clarifications documents since they had only recently been made available. However, we were struck and particularly disturbed by the fact that only half of the teachers used the exemplars, since of all the state-developed materials, these were intended as staff development tools to model effective performance-based lessons and were not test-oriented. Although we did not ask teachers about their familiarity with and use of other resources (e.g., commercially produced tasks or those created by educators from other systems, states, or by the Maryland Assessment

Consortium[3]), later examination of instructional and classroom assessment materials would reveal that teachers are often aware of, sometimes using, and often misusing, a wide array of materials "marketed" as performance-oriented.

In addition to identifying their knowledge and use of existing resources, teachers also identified other information and support that they would find most helpful. While several teachers used this questionnaire item as an opportunity to identify such needs as smaller class size, more preparation time, or greater community involvement as critical to their work, by far the resource most frequently requested (by more than half the teachers surveyed) was a larger pool of practice tasks across content areas that would be related both to the MLOs and specifically to the Charles County Curriculum framework. Five additional teachers similarly requested more public release tasks of better quality than some of the "retired" tasks that had already been made available by MSDE. Several other teachers requested more help in developing tasks and rubrics, asking for more staff development under the guidance of a consultant or specialist (rather than other teachers) or even a newsletter that might offer general guidelines and "hot tips."

Both at the beginning of scoring and after scoring, teachers were asked to identify selected topics that they would like to know more about. Overall, teachers expressed the most interest in learning more about helping students with self-assessment (19, 25), different strategies for judgment-based scoring (18, 22), the relationship between the MLOs and the county curriculum framework (22, 21), what makes a task "scorable" (21, 18), essential

---

[3] The Maryland Assessment Consortium is a collaborative representing the majority of Maryland's 24 school systems devoted to creating and distributing formative assessment tasks intended to measure the MLOs but not strictly modeled along MSPAP.

characteristics of performance tasks (16, 18), and performance task design (17, 15). After scoring, teachers expressed somewhat less interest than before in knowing more about what makes a task "scorable" (perhaps because this was explicitly discussed during operational scoring) but significantly more interest in knowing more about helping students to develop better self-assessment skills and communicating to students their proficiency in the MLOs, as well as in developing different strategies for judgment-based scoring. They also expressed somewhat more interest in learning more about communication with parents and the essential characteristics of performance assessment tasks and activities. These patterns were confirmed during the interviews in a variety of ways. Perhaps most important, many of the teachers interviewed stressed the importance they placed, thanks to the scoring experience, on developing students' self-assessment strategies and, more generally, on developing students' independence as learners. In their references to "the big picture" and the potential of MSPAP to foster improved learning, they also repeatedly stressed the importance of communication about the goals and meaning of MSPAP with multiple constituencies.

Data from interviews

After analyzing the questionnaire data and considering some of the materials teachers had shared with us, we conducted a series of twelve interviews with teacher-scorers in order to gain more insight into teachers' perceptions and priorities. The interviews gave us an opportunity to ask more directly about teachers' responses to the scoring experience now that they had been back in their classrooms for a semester.

When asked about the value of scoring, every single teacher interviewed responded with some version of three comments:

1) Scoring was such a valuable experience that it would be ideal if every teacher and administrator could score.

Some teachers suggested that scoring would be a valuable component of pre-service education for teachers, and several said they learned more from scoring than from most education courses they had taken. It is worth noting that since the earliest years of MSPAP, teachers have given similar testimonials to the value of the scoring experience, expressing the wish that "all teachers ...become involved in scoring the MSPAP" (Goldberg, 1994). Also interesting was that several veteran scorers reported that because the scoring experience is itself so intense (one likened it to childbirth!), they gained more insights applicable to the classroom after the second year of scoring.

2) Scoring gives you the "big picture" and serves as a "wake up call."

Most teachers indicated that it was very valuable to step beyond the isolation and idiosyncracies of an individual classroom or group of students to see the range of possibilities of student work. A major consequence of the "big picture" was that teachers were galvanized by what they saw in scoring to raise expectations, either because they saw what some students could achieve, or because they saw the dangerous consequences of failing to expect the most of students. Teachers reported that they saw their own teaching much more clearly as part of a larger ongoing educational process, and left scoring feeling more accountable for their role in this larger process. One teacher, concluding that scoring made him "more ruthless, but more liberal" summarized well the perception many teachers had that scoring led them simultaneously to raise standards and to be more flexible in allowing for different ways to meet those standards.

3) Scoring "makes you think."

Many teachers reported that scoring had made them more critical and deliberate in their work by inviting them to more carefully scrutinize tasks, student responses, and the criteria by which responses are evaluated. While teachers divided evenly between those who emphasized the impact on their instructional practices and those who said what was most changed were their evaluation and assessment practices, each teacher spoke of becoming more thoughtful and more focused on determining the goals of his or her teaching and assessment and how these aligned with the Maryland Learning Outcomes.

Overall, the interviews confirmed and added emphasis to many of the responses to the questionnaires we distributed. Every teacher interviewed believed that MSPAP either had improved or had the potential to improve teaching and learning, largely by encouraging the use of more hands-on and integrated activities, including more reading and writing of various types and for various purposes, emphasizing the importance of higher order thinking, explanations and text support, and raising expectations.

Teachers' interview comments also powerfully demonstrated that the pathway from scoring to classroom practice is neither direct and linear nor simple and predictable. Instead, their comments revealed that the scoring experience is mediated by a variety of factors, including not only teachers' past experiences and personal approaches to teaching, but also school and county-wide directives concerning the tests and best ways to improve teaching and learning, the different ways the test is defined and "packaged" in schools and other instructional communities, opportunities for collaboration with other teachers, and ongoing staff development supporting performance-based instruction.

The interviews pointed to the especially productive role many resource teachers were

able to assume following training. Those teachers who staffed resource rooms, regularly visited multiple classrooms, or served as grade leaders were both more likely to come into close and sustained contact with other teachers around instructional issues and, more importantly, were already in a consultative role that made sharing their expertise comfortable. While some classroom teachers indicated in the interviews--and in the questionnaire--that they were afraid to "push" or that they did not believe their suggestions or insights would be welcome, resource teachers and team members spoke without ambivalence about sharing materials they had created and more generally "spreading the word"; one media resource teacher said she was "spending fully half [her] time reviewing tasks" for other teachers, while one language arts specialist estimated that 95-99% of her time in third, fourth and fifth grade classrooms was focused on improving scores on MSPAP, working directly with students and teachers on tasks aligned with MSPAP. These comments are indicative of the fact that while many teacher-scorers had been authorized to, and had indeed assumed significant responsibilities for helping other teachers to implement performance-oriented activities, their efforts were much more often focused on MSPAP per se than on performance-based instructional and classroom assessment strategies.

Clearly, different school administrators have very different approaches to the demands and challenges created by MSPAP. Very few members of the administration have scored the test, and teachers often commented that not only should all teachers score, but all administrators should as well. Both the reassignment of teachers with scoring experience to grades 3, 5 and 8 and the requirement that teachers post MSPAP rubrics, descriptions of the purposes for writing, and other "canned" documents on the walls of classrooms, pointed to the ways that school administrators were often much less reflective, though no less anxious, about ways to prepare

students for the tests. These documents, never intended for display purposes and some never intended for students' use at all, often functioned to create a kind of "noise" in the classrooms, emblematic of the ways that administrators, and in turn teachers, seemed to hope that continual exposure to MSPAP rubrics, scoring tools and content area descriptions would somehow infuse students' learning and obviate the need for more dramatic and sustained scrutiny and revision of what was being taught and learned and how. This focus on a "quick fix," and the implied expectation that student scores show noticeable improvement over the previous year's, generally served to create tension for teachers without providing real support for the kinds of curricular and instructional innovation that would lead to improved learning. Some schools' decisions to offer McDonalds food, provide candy, or sponsor dances and special events as a reward for participation in MSPAP lend further credence to the notion that school leadership may place more priority on raising scores than on sponsoring meaningful educational change. We must acknowledge, however, that there is great pressure to do so because of the ever-increasing threat of state "reconstitution" of inadequately performing schools and the current system of sanctions and fiscal rewards which operate in Maryland.

In some schools, the desire to institute more performance-based instruction translated into what one teacher said had become the eleventh commandment: "Thou shalt create tasks." It was this mandate that teachers create multiple "mega-tasks" -- tasks that would approximate the longer integrated tasks in MSPAP and would assess multiple content areas through the use of manipulatives and other hands-on activities -- that created the most resistance on the part of even those teachers who were interested in enhancing performance-based instruction in their classrooms. One primary grade teacher, for example, said the impact of MSPAP on teaching in

his school "could be summed up in four words: work, work, work, work." The resource teacher who spent "half [her] time" reviewing tasks for other teachers in the school similarly complained that teachers had been given the message that the quality of their teaching could be measured by the number of mega-tasks they designed. In these respects, the administrations' partial and often superficial, "what counts is what you can count" attitude paralleled the more superficial approximations of MSPAP-like tasks which may be seen in materials created by teachers least familiar with the test.

It is important to note that teachers' resistance to the demand that they create tasks and their persistent and universal requests that "more tasks" be provided to them was not a simple matter of lack of time or energy. Instead, several teachers spoke of feeling "overwhelmed" by what they saw as an inappropriate demand that they, individually, essentially become test developers and create and field test complex tasks and scoring criteria that would integrate science with other content areas. This sense of being overwhelmed was heightened, in part, by the tensions between the county curriculum documents and the Maryland Learning Outcomes. This was especially true in schools that had also adopted other ambitious, cross-curricular initiatives like the "Going Places" program that introduced yet another distinct agenda into the already overburdened and sometimes contradictory curriculum. One of the teachers who had been most successful in using her scoring experience to rethink her teaching in productive ways further clarified this problem through the traditional distinction made between curriculum and instruction. Believing that her major role was to focus on instruction -- how to present activities and information, how to tailor material to a large class with a wide range of abilities, how to address individual students' needs -- she complained that the emphasis on task creation made

teachers responsible for curriculum development at the expense of instruction, and argued strongly for the need if not for ready-made tasks, then for more and better models, a better library of appropriate resource materials for teachers to draw on, and more professional support for creating curricular materials.

In multiple ways teachers' comments in interviews revealed that the emphasis on performance-based instruction had been layered on top of an existing curriculum, rather than inviting a rethinking of that curriculum. This was how the administration communicated its expectations, and it was also how all but the most experienced scorers attempted to include performance-based instruction in their teaching. Teachers voiced concerns about losing content to the test, indicating that they saw the emphasis on higher order concerns and integration as antithetical to -- rather than supportive of -- the learning of "content." Several teachers spoke of the difficulties of keeping accurate records when doing performance-based instruction, rather than considering that performance-based instruction and assessment might also require a re-examination of student evaluation and the ways that grades are assigned and recorded. Teachers at several schools spoke of the continual pressure to "dream something up" that would look like one of these mega-tasks, a phrase that reveals their assumptions that these tasks will necessarily be contrived. Like the questionnaire responses in which teachers indicated their plans to "take literature and try and put more math and science into it" or to "do MSPAP daily or weekly," the interviews revealed that teachers often saw content integration or performance-based activities as a matter of "tacking something on" to an existing lesson or topic. Trying to explain why this general perception persisted, one teacher who had participated in scoring for three years lamented that MSPAP is presented not as a model for a way of thinking about teaching and

learning but as a distinct object or artifact, moving erasers around on a table to demonstrate how knowledge about the test is conceived and communicated.

What Sample Activities Reveal

Like the questionnaire data and interviews, the instructional and classroom assessment activities we gathered demonstrated the complex ways that teachers apply their understandings based on MSPAP to their classrooms by highlighting the differences between the perceptions and practices of teachers who have had scoring experience and those who have not. Sample instructional and classroom assessment activities developed by teacher-scorers shared various characteristics that were absent or less evident in materials developed by their colleagues. These characteristics include attempts to: 1) establish context and purpose; 2) align activities with MLOs and indicators; 3) include opportunities to read and write for a variety of purposes and audiences; 4) provide content integration; and 5) formulate and use evaluative criteria.

Establishing context and purpose

One of the key characteristics of MSPAP tasks is that they are based on plausible, real-life situations, problems, issues, or decisions, and are comprised of a series of activities for which the purposes are clear and authentic. Because MSPAP is a paper-and-pencil test, only constructed responses such as a piece of writing, a drawing, diagram, or graphic display of some sort can provide a measure of proficiency in one or more outcome areas. Therefore, a typical purpose for doing a series of activities might be to gather information to allow students to make an informed interpretation, recommendation or plan, to be communicated through a report, speech, or data display intended for a clearly identified audience.

Typically, the lessons developed by teachers lacking in scoring experience demonstrate

at best limited efforts to establish a context and purpose beyond that of "academic exercise." Even when activities are joined by a common theme ("Japan," or "Native Americans," for example), teachers do not clearly establish for students some real-world reason for what they will be learning and doing. Students are not provided with a sense of where what they'll be doing is leading, or how they can expect to apply what they are learning. The "M & M task," a set of instructional activities which has been frequently and variously modified in different primary grade classrooms around the state, is representative. In one Charles County version, students tally the number of M&Ms of each color that they find in a single-serving packet; they then do some basic computation (M&M math) and then complete two writing "starters" (see Figures 1-3). While this set of activities is undoubtedly engaging (especially since students can eat the manipulatives at the end of the lesson), the rationale for this series of activities is left unstated and is merely a curricular one--to teach graphing (statistics), review computation, and give students an opportunity to write. Students are never told or led to discover for themselves any connections beyond the thematic one, nor do they ever consider what they might do with what they have learned. This set of activities is particularly interesting because on the surface it looks like an ideal one--it is engaging, incorporates the use of manipulatives, involves cooperative learning, and draws on knowledge in different domains. It fails, however, to involve students in solving a real-life problem, marshaling what they know and can do in order to achieve a goal.

In contrast, teachers with scoring experience tend to create lessons/units with at least a rudimentary and somewhat coherent framing of context. For example, in one performance-based lesson, the conflict between the tobacco industry and the medical community becomes the

context for a series of reading (to be informed), social studies (economics), and mathematics (statistics) activities that culminate in students using both their own ideas and the information they have gathered from a variety of resources to write a letter to persuade the President to support their position on a proposed law that would make cigarettes illegal. While this lesson might have benefited from more preliminary discussion of the tensions between the County's long-standing economic base in tobacco farming and students' personal concerns for the health of family members, it successfully establishes a believable and compelling context and purpose.

Aligning activities with Maryland Learning Outcomes and indicators

Prior to the inception of MSPAP, teachers in Maryland were guided by curriculum framework documents developed by each local jurisdiction. With the formulation of the M.O.'s as a step towards the development of MSPAP, local educational agencies were pressed to review and revise these frameworks to ensure that the MLOs were addressed and that by following county curriculum, teachers could rest assured that students would be well prepared to demonstrate proficiency in the areas assessed by MSPAP.

Since the Charles County curricular framework document was revised in 1996, teachers have been told not to worry because the MLOs are "in there." The wide-spread assumption among teachers, therefore, is that if they follow the framework, the learning outcomes will somehow all be addressed. Lesson plans typically come adorned (for principal's scrutiny, no doubt, rather than for any real pedagogically valued reason) with a listing of Charles County "targets and indicators" embodied therein. Although familiar with the Charles County framework, most teachers can at best name the content areas assessed on MSPAP and are unfamiliar with the precise indicators of proficiency in those areas, even though test items and

scoring tool criteria are developed based on descriptions of M.O. indicators. For example, while teachers know they need to cover the outcome, geography, they are unlikely to identify the ability to locate information on a map as only the first of over a half dozen indicators for that outcome. The use of the term, "indicator," as both a sub-set of county curricular targets and state learning goals is confusing to teachers and interferes with their understanding of the construct underlying MSPAP and its relation to instruction.

This confusion may contribute the fact that as of yet, there have been almost no attempts at curriculum mapping (cf. Jacobs, 1997) based on the MLOs and indicators. This seems to have led to a situation not uncommon at family picnics, where all those assembled suddenly stop to inquire, "Who has the pickles?" and discover that in the absence of communication about what is expected, and from whom, there are a dozen tubs of cole slaw but no pickles.

Because of their lack of familiarity with the range and detail of the MLOs and indicators, many teachers are generating instructional and classroom assessment activities that are characterized by what little they do know about MSPAP. These classroom activities are sometimes poorly aligned with what is actually assessed and may do little to prepare students for the test. With limited exposure to MSPAP, as their questionnaire responses showed, teachers tend to think of performance-based activities as constructed response, hands-on, collaborative, and open-ended, without recognizing that these features are a means to learning the skills, processes, and knowledge encompassed in the MLOs rather than learning goals in themselves. This has led to the proliferation of "mini-MSPAPs" which have the appearance but not the substance of good performance-based instruction or assessment.

A good example of an "empty" activity is one which, ironically, the teacher who crafted

it called "MSPAP Activity" (see Figure 4). In this open-ended activity, students are asked to decide upon, and then work collaboratively to craft an item to add to the interior of a clubhouse. Other than providing a scenario, a list of available materials (boxes, scissors, glue, staplers, tubes, etc.), and instructions to "work cooperatively and have fun!" students are left without any sense of what skills or strategies they might wish to (or are in fact required to) employ. This activity has the potential to provide an opportunity to teach problem solving, measurement, and estimation (mathematics), and the concept of the relationship between available resources and the production of goods (social studies/economics), if modified to prompt students to work within articulated parameters and to address certain steps or questions as part of the task. As is, however, students may wind up happily, busily engaged--in nothing that will ultimately lead to greater proficiency in the outcomes that are supposed to underlie instruction and assessment.

Even among teachers who have scored MSPAP, there appears to be confusion between the opportunity to address a given outcome and an occasion either to teach concepts and processes related to that outcome or to obtain a measure of proficiency in that outcome. Thus, for example, a host of opportunities are lost in a performance task which springboards off of the reading of Jumanji, by Chris Van Allsburg. Students respond to a series of questions about this novel, which deals with a board game gone out of control. Then, after brainstorming other board games they know of and have played, students work in groups to create a new game using one or more of a set of objects provided (e.g., drinking straw, marble, marker, macaroni, metal washer). Students play their game and those of other groups, evaluating each game in terms of whether or not it was fun to play, similarities and differences among games, and ideas about things to change in each game. By observing "tournament" competition, students next identify and

resolve problems they observe teams having as they play the newly invented games, and recraft instructions. Finally, students are asked to write an advertisement to try to persuade people to buy the new game they've created. The teacher who created this task identified at the outset the MLOs and Indicators being addressed through these activities including, for example, political systems (describe the processes people use for making and changing rules within the family, school and community) and understandings and attitudes in social studies (propose rules that promote order and fairness in various situations); nevertheless, while the scenario of creating and evaluating games might have been effectively employed to develop understanding of these processes and concepts, the promise of this task is unrealized in terms of both instruction and assessment. Although students are led through many things that are certainly worthwhile, the modeling on MSPAP does not appear to have had any meaningful consequences in terms of teaching and learning the intended social studies outcomes and indicators.

This particular task also illustrates the need for staff development support in another regard--understanding the need for instructional and assessment activities to pertain to some "overarching" idea in order to provide coherence to performance-based lessons and tasks. The reading questions on Jumanji were of the sort teachers typically ask--some involving simple information location and retrieval, some involving interpretation and inference, and yet another, a "personal reflection" question that in fact did not cause students in any way to reflect back on, or construct, extend, or examine meaning in the literary selection. Innocuous enough as questions go, they nevertheless squandered the chance to use the reading selection as an entry-point to considering the concepts which the task was intended to address--how and why rules are made by groups of people. With some revision, students' reading for literary experience might

have provided for meaningful consideration of rules and instructions in "the games people play."

Including opportunities to read and write for a variety of purposes and audiences

MSPAP measures students' ability to read for three purposes--for literary experience, to be informed, and to perform a task--and to write to inform, to persuade, and to express personal ideas. Teachers who have scored MSPAP seem generally more familiar with the reading and writing construct, and build in more opportunities to read and write for a variety of purposes. Nevertheless, even among this more highly informed population, certain misconceptions and omissions in practice prevail.

Across purposes for reading, even teachers with scoring experience struggle to craft the range and variety of "stance" questions (see Langer, 1989, 1990; National Assessment Governing Board, 1992) that guide students' orientation to the text--as they read for global understanding, to develop interpretation, to formulate a reader-text connection (personal stance) and/or a critical stance (by considering not what, but how, meaning is made). Reading questions continue to mirror textbook-style, lower level reading skills, and to encourage information location and retrieval, a process of "reading with one's finger." County-wide, the vast majority of reading activities center on literary texts. Informative selections are far less common, and those which enable the reader to follow directions or conduct an investigation are rare indeed. Even among the "cognoscenti" who have scored, and are more likely to provide classroom opportunities to read "perform a task" selections, reading activities often entail no more than first reading, and then immediately doing, an activity. There is little or no discussion to guide students through the construction, extension, and examination of meaning that must occur when students interact with this type of text as with any other.

Writing and language usage are the only areas scored with generic criteria, or rubrics, which because they are not activity-specific are not secure.[4] Teachers are generally familiar with the purposes for writing assessed on MSPAP, and they often have the rubrics posted on their classroom walls (even, we noted, in developmentally inappropriate contexts). Once again, however, being able to name outcomes is not evidently the same as understanding how to teach to them, or measure student proficiency in them. Across grades, teachers with and without scoring experience cue their students to write to inform, to persuade or to express personal ideas. However, even among teachers who scored, a tendency to cross-cue prevails. So, for example, students might be asked to "imagine" that they held a certain job and then "inform" others about that job. Once students' "creative writing" button has been pushed with the cue to "imagine," even the explicit cuing "to inform" may not keep them from drifting from marshaling and organizing plausible ideas and information to increase a reader's understanding of a topic. Similarly, even after cuing students to "persuade the principal to buy new playground equipment," a series of informational "think abouts" may cause many students detour from the intended purpose (see Figures 5 and 6). Since scoring is purpose-specific, such writing activities do not familiarize students with key characteristics of writing for the purposes ultimately to be measured or serve them well in developing awareness of different strategies that might be employed for varying purposes and contexts.

Providing content integration

Another of the design features of MSPAP generally familiar to teachers (both with and without scoring experience) is the integration in many tasks of activities that address outcomes in different content areas. For many teachers, however, this awareness has been delivered

through a system and school leadership mandate to "integrate," unaccompanied by any staff development on the ways and means of doing so. Typically, among teachers without scoring experience, content integration takes only the most superficial form of activities addressing different content areas "in tandem"--for example, a set of reading activities followed by a set of science activities, followed by a writing activity. The loose thematic umbrella described earlier often becomes the mechanism for including multiple content areas, although no effort is made to build student understanding either of commonalities among content areas or of the different conventions sometimes associated with particular disciplines.

Content integration is perhaps the feature which has been most often internalized, and with the greatest success, by teachers who have scored MSPAP, and is again linked to facility in context setting. As teacher-scorers establish real-world contexts for a unit of investigation or exploration, they tend to weave in activities that cut across a variety of content areas in an uncontrived way. In one integrated unit, based on reading a chapter in a book about early Americana, students complete a graphic organizer on colonial inventions, chose one, and reflect on why it was invented and how it helped colonists meet their needs and wants. Students consider the impact of available resources on the production of various inventions, and then plan a way to construct a model of one invention for a school-based colonial fair. Reading, writing, and social studies (both economics and peoples of the nation and world) weave smoothly through this set of activities. Rather than merely adding on or providing a series of takes or snapshots, understanding is augmented by examining concepts through the kaleidoscopic lens of multiple content areas.

Formulating and using evaluative criteria

Given that teachers participating in this study were exposed not to task development, but to scoring, classroom assessment strategies are perhaps the area in which one would most expect to see some direct and positive impact from the scoring experience. Indeed, teacher-scorers seem generally to understand the physical format for designing criteria (e.g., the "stepping-stone" framework in which different degrees of evidence of various characteristics, rather than different characteristics, define each score point). Yet, exposure to the use of evaluative criteria directly linked to the outcomes and indicators appears to have had many unanticipated consequences. The most positive effect of scoring has been the adoption by many teacher-scorers of the MSPAP writing and LU rubrics and rules (the condensed version used for brief constructed responses scored for these areas) in the classroom. Even in this regard, however, there is some confusion, with a number of teachers using the 0-3 scale reserved on MSPAP for extended writing activities (those in which students employ writing process strategies to develop their work) to score brief constructed responses. Teachers who have scored MSPAP are virtually alone in understanding that the activity-specific keys used to obtain all measures on MSPAP except those in writing and LU are crafted using the language of the MLOs/indicators, and that there must be alignment between what is taught and how what is taught is evaluated. However, this understanding is not translating well, as of yet, into practice.

While teachers who have scored MSPAP are, far more often than other teachers, crafting and using evaluative criteria, these criteria often demonstrate one or more flaws. These include confounding the outcomes being measured, scoring for extraneous features (e.g., neatness, color, etc.), scoring by counting up parts or components rather than looking for evidence of proficiency in the outcome(s) being measured, scoring for things they have not cued students to do, and

scoring products rather than outcomes.

The flaw most often observed can be described as confounding of outcomes. Within a single scoring tool, criteria for multiple outcomes are merged under score point descriptors, such that the same level of performance is expected to characterize novice, intermediate, proficient, and expert level regardless of content area skills and processes being demonstrated (see Figure 7 for example). In actual practice, it is far more likely to see evidence of differing degrees of proficiency in, for example, reading, writing, and language usage skills such that a student might be performing at a 4-level in reading, a 2 in writing, and only a 1 in language usage. Confounding of outcome descriptors causes whoever is making a score decision to compromise and often "settle" on a midrange score, thus providing a measure that is not valid for any outcomes being assessed.

One of the axioms of scoring performance assessment which has been widely shared in Maryland is that "you don't score by counting on your fingers" (Goldberg, 1995). Often, when teachers have had initial but limited exposure to "rubrics," they translate the framework of score point descriptors into the most mechanical of schemas, whereby "four examples" yields a 4, "three examples" yields a 3, and so on. There may be little or no thought given to whether the quantity of ideas, examples, reasons, etc. is valid evidence of proficiency in the outcome or indicator being assessed. In fact, unless those crafting scoring criteria can provide a logical and convincing rationale for cuing score decisions with counts, these should not be a feature of scoring tools. Nevertheless, scoring tools flawed in this way abound (see Figure 8 for example).

Even when teachers successfully cluster performance characteristics by outcome, teacher-crafted tools often include extraneous features. While purportedly measuring

performance in the MLOs, scoring criteria often include descriptors better categorized under work habits or creative expression (see Figure 9). Although there is certainly no injunction against measuring these traits, the same guidelines for creating effective scoring tools must apply--if students can demonstrate differing degrees of proficiency in various areas, they require separate scoring tools; furthermore, care must be taken not to taint data on MLO performance with information more correctly subsumed under other instructional objectives. These same "grading criteria" in Figure 9 also demonstrate the tendency of many teachers to score for completion of a product rather than for evidence of proficiency embodied in that product. When evaluative criteria are linked only to the specific demands of a given task or activity, however (like making a puppet), they are, as W. James Popham (1997) recently noted, "essentially worthless."

An even more serious variant of the problem of attending to extraneous features is evident in scoring tools which include descriptors for uncued-for features of the response being evaluated. Some teachers' tendency to reward credit for "something extra," or that "je ne sais quoi" intangible quality that somehow "elevates" products and performances finds its expression in the assignment of higher score points only to responses which serendipitously demonstrate a feature the need for or desirability of which was never made clear to students.

Ironically, it is often in those scoring tools that are most explicit in their association in teachers' minds with MSPAP that the most egregious distortions of valid judgment-based scoring occur. Thus, a "Rubric for MSPAP Activity" (see Figure 10), for example, illustrates confounding of outcomes (problem solving, computation, writing, language usage), scoring for extraneous elements (details and color), scoring for things students were not cued to do (adding

and coloring an illustration), and counting on fingers (paragraph with 5-7 sentences). Furthermore, this scoring tool illustrates a "Chinese menu" approach to evaluation whereby a student may receive the same score for doing entirely different things--scores which are meaningless because not aligned with any given content area and "homeless" measures because teachers cannot find any place in their records of student progress to capture this information.

Conclusions

Almost without exception, teachers endorse the scoring experience as a valuable one which galvanizes them, and makes them more reflective, critical and deliberate. Thanks largely to their scoring experience, their classroom activities are more likely than their colleagues' to elicit writing for varied and coherent purposes, to integrate content, and to cue for higher order thinking. At the same time, however, like Socrates' wise man who knows that he does not know all, teachers report that the experience highlights for them the as yet unfulfilled need for resources and professional support in order to meet demands and expectations that only grow greater and more complex with their increased understanding of the issues and implications of performance-based instruction and classroom assessment.

While the scoring experience often challenges and energizes teachers, it does not provide them with comprehensive and well-grounded understanding of performance-based instruction. This study suggests, instead, that the scoring experience does not automatically or easily translate into effective classroom practice. Although judgment-based scoring is more and more frequently being touted as a powerful opportunity for staff development, we find that the experience of judgment-based scoring, by itself, is likely to yield limited benefits. Even in schools with faculty who have been trained to score MSPAP tasks and have participated in

locally-designed staff development on the scoring process, the impact of exposure to scoring tools and methodology has still been limited, and teacher-generated activities typically:

- Are often interesting and engaging, but bear little or no connection to the MLOs and/or indicators; where such connections are articulated, they are identified only at lesson or task level--not at the activity level
- Are preceded by little or no context-setting, whether used for instructional or assessment purposes; at best they are a series of activities with a thematic or topical connection
- Often cue for skills and understandings extraneous to their intended purposes
- Have been transformed into "worksheets" even when intended as organizers (webs, story maps, etc.)
- Cue for recall and information-location rather than higher-order skills and processes modeled in MSPAP

Classroom assessment strategies tend to show even less evidence of any positive impact from exposure to the application of MSPAP scoring tools and scoring methodology. In general:

- Teacher-developed tasks confuse the opportunity to see evidence of a given outcome with the conditions under which it may be measured
- Learning outcomes are frequently confounded in scoring tools so that one set of criteria is intended to provide information on different areas in which students commonly demonstrate varying degrees of proficiency; sometimes scoring criteria do not even address any of the MLOs
- Score point descriptors are often arbitrary and trivialize what is being measured by focusing on what is easy to count up or pluck out

- Rather than encouraging responses which require the creation of some product (e.g., constructed responses, drawings, schematics, graphs and charts) tests reflect traditional item types such as matching, fill in the blanks (often from word banks) or true/false; even performance-oriented instructional activities wind up being graded and embellished with check marks, percent right, or a "smiley face"

Educational Importance

In the absence of a clearly articulated and well-disseminated rationale for performance-based instruction and assessment, supported by sustained professional development and the services of state and local specialists to help accomplish curricular goals, many teachers have struggled valiantly in approximating the kinds of instruction that programs like MSPAP are intended to foster. With few models and limited support or opportunities for collaboration, they have gone about the business of dissecting tasks, translating abstract outcomes into teachable lessons, and transforming a complex performance assessment model into classroom practice. That their approximations should themselves be partial or imperfect should come as no surprise.

By themselves, neither summative, state-mandated assessments nor the opportunity to participate in evaluation of students' work are likely to create the desired differences in teacher thinking and practice envisioned in school reform. Indeed, the assumption that even the best state-wide performance assessment can directly model and improve instruction and learning has itself proven overly simplistic by the teachers who shared their ideas and materials with us. However, their comments are invaluable in highlighting what teachers believe and need, just as their classroom materials provide detailed evidence of which concepts are most easily appropriated and internalized, and which are most elusive.

The need to go beyond anecdotal accounts of the benefits of the scoring experience to determine what additional supports are needed becomes ever more critical as new assessments are under development, both in Maryland and elsewhere, which include plans for all judgment-based scoring to be done by teachers and at a local level. Even more pressing, at a time when national testing is under consideration, is the recognition that tests alone will accomplish few of their goals without sustained and multi-layered staff development that builds upon what teachers already understand and are doing to help students learn to apply knowledge meaningfully in a performance-oriented context.

We would be well-advised to also recall, and endeavor to hold true to, the vision that first led to the creation of MSPAP and presumably underlies other large-scale assessment programs. That vision is one of a program that drives school and instructional improvement and models exemplary teaching and learning, while providing valid and meaningful accountability data (Sondheim, et.al, 1989). As steps are undertaken to provide comprehensive accountability systems and state-of-the-art data management, we must not lose sight of the need to support instructional improvement initiatives and a system of timely and thorough dissemination of resources and assistance to school level personnel. This is imperative even in the face of staff and budgetary limitations (Hettleman, 1998). Such responsibilities as the identification and publication of information on exemplary programs, practices, and staff training models and the establishment and maintenance of training centers and highly skilled trainers must not be neglected. The wisdom of the initial, more comprehensive vision behind MSPAP is mirrored in the works and actions of the teachers to whom we listened and from whom we learned.

References

Afflerbach, P., Guthrie, J., Shafer, W., & Almasi, J. (1994, April). *Barriers to the Implementation of a statewide performance program: School personnel perspectives.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Goldberg, G. L. (1994). Learning the score: What teachers discover from scoring performance assessment tasks. *Teaching Thinking and Problem Solving*, 16 (1), 1, 3-6.

Goldberg, G. L. (1995). Gail's axioms of scoring performance assessment. In Neill, M. et al, *Implementing performance assessments: A guide to classroom, school and system reform.* Cambridge, MA: Fairtest.

Hettleman, K. R. (1998, January 18). Looking past Md.'s top marks. *Baltimore Sun*, 1G, 8G.

Koretz, D., Mitchell, K., Barron, S.& Keith, S. (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Jacobs, H. H. (1997). *Mapping the big picture: Integrating curriculum and assessment K-12.* Alexandria, VA: ASCD.

Langer, J. A. (1989). *The process of understanding literature* (Report Series 2.1). Albany, NY: Center for the Learning and Teaching of Literature.

Langer, J. A. (1990). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English*, 24, 229-260.

Moss, P. (1997, March). Untitled presentation. In W. Yen (Chair), *Investigating the consequential aspects of validity: Who is responsible and what should they do?*

Symposium conducted at the annual meeting of the American Educational Research Association, Chicago.

National Assessment Governing Board. (1992). *Reading framework for the 1992 National Assessment of Educational Progress.* Washington, DC: Author.

Popham, W. J. (1997). What's wrong--and what's right--with rubrics. *Educational Leadership* 55, 2 (October 1997): 72-75.

Sondheim, W. Jr., et. al. (1989). *The report of the governor's commission on school performance.* Annapolis MD: Authors.

Waldron, T. MSPAP gets okay of most teachers (1997, July 16). *Baltimore Sun*, pp. B1, 3.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership* 49, 8 (May 1992): 26-33.

Appendix A:

Interview Questions: What has been the impact of scoring MSPAP on your teaching?

- What are your more general perspectives and feelings about MSPAP? What has been the impact of MSPAP on your student's learning? On your teaching?
- Can you point to any ways that helping students prepare for MSPAP has improved your teaching?
- What motivated you to score MSPAP?
- What do you see as the value of the experience of scoring MSPAP?
- What, if anything, have you done differently or will you do differently in your classroom a result of participating in scoring?
- Has MSPAP scoring had any effect on the ways you evaluate student work?
- (Do you use rubrics or other scoring tools in any way in your teaching? Why or why not? With what effects?)
- Before and apart from scoring MSPAP, what was your experience with performance based instruction and performance assessment? And after?
- What do you see as the key elements of establishing and maintaining performance based instruction and assessment ?
- What are the main challenges in doing performance based instruction?
- What kinds of resources and support would be most valuable to you in creating a performance-oriented classroom? In preparing your students for MSPAP?
- In what ways has your expertise been used and shared?

- Would you be willing to give me any examples of teaching materials you created before and after scoring that show these differences?

Table 1
Comparison of First Year and Veteran Scorer Ratings of Their Own and Colleagues' Knowledge of Performance-based Instruction and Performance Assessment

Testing Experience		Knowledge prior to scoring	Knowledge after scoring	Teachers' knowledge	Administrators' knowledge	Knowledge of performance assessment before scoring	Knowledge of performance assessment after scoring
first year	Mean	2.7500	3.4737	2.6250	3.0000	2.7500	3.5500
	N	20	19	16	15	20	20
	Std. Deviation	.7164	.6118	.6191	.3780	.8507	.6048
more than one year	Mean	2.8824	3.7059	3.0625	3.0000	2.5882	3.5882
	N	17	17	16	15	17	17
	Std. Deviation	.7812	.4897	.4425	.8452	.7952	.5073
Total	Mean	2.8108	3.5833	2.8438	3.0000	2.6757	3.5676
	N	37	36	32	30	37	37
	Std. Deviation	.7393	.5542	.5741	.6433	.8183	.5548

Testing Experience		Teachers' knowledge of performance assessment	Administrators' knowledge of performance assessment	Use of performance based instructional activities before scoring	Use of performance based activities after scoring	Use of performance assessment prior to scoring	Use of performance assessment after scoring
first year	Mean	2.6875	3.0667	2.8333	3.6000	2.8889	3.5500
	N	16	15	18	20	18	20
	Std. Deviation	.7042	.4577	.7071	.5026	.5830	.5104
more than one year	Mean	2.8750	3.0000	2.7647	3.7647	2.5294	3.7647
	N	16	15	17	17	17	17
	Std. Deviation	.6191	.9258	.7524	.4372	.6243	.4372
Total	Mean	2.7812	3.0333	2.8000	3.6757	2.7143	3.6486
	N	32	30	35	37	35	37
	Std. Deviation	.6591	.7184	.7195	.4746	.6217	.4840

BEST COPY AVAILABLE

Figure 1:

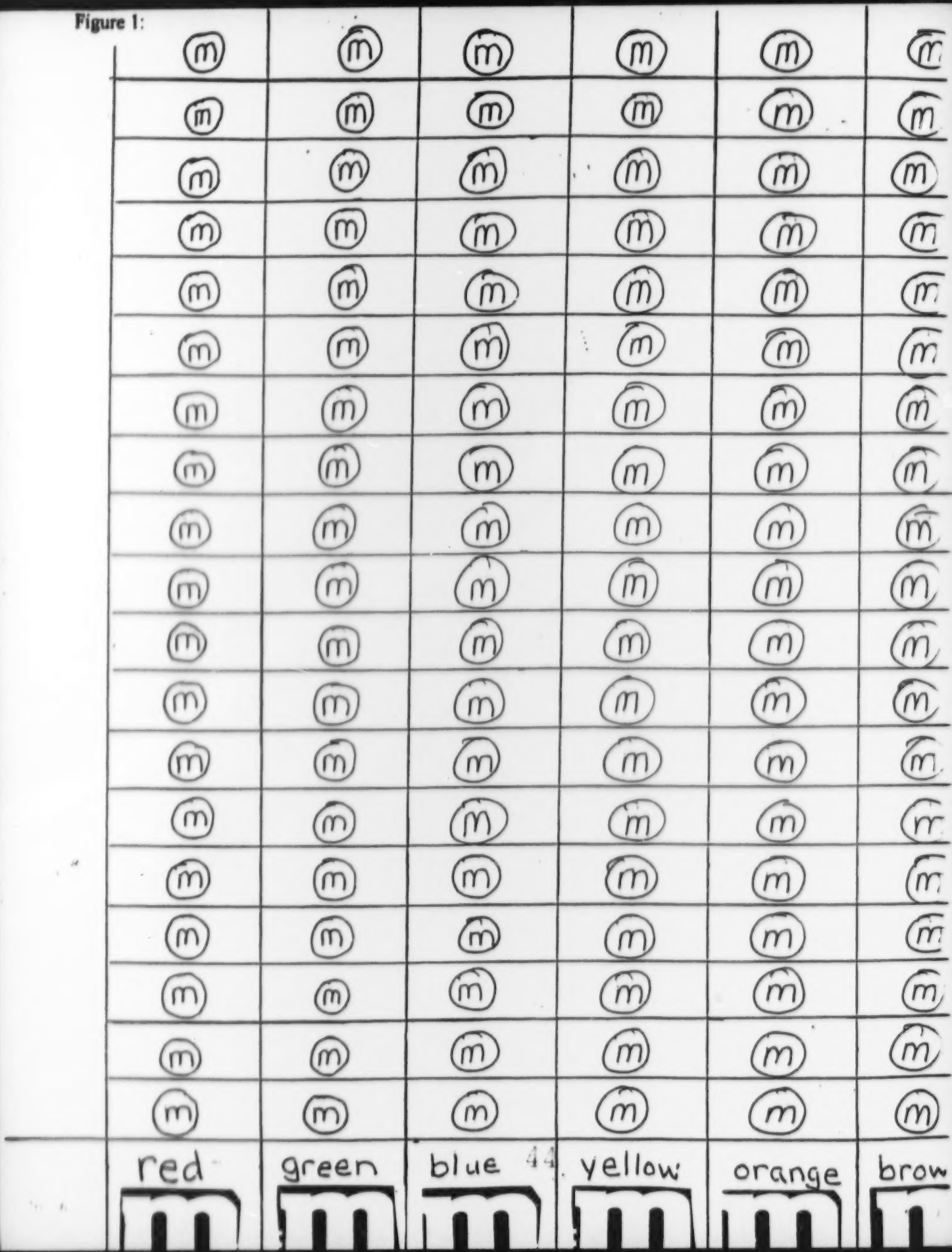

Figure 2.

Name _____ M&M addition

_____ + _____ = _____ M&M's
red blue

_____ + _____ = _____ M&M's
green yellow

_____ + _____ = _____ M&M's
orange brown

_____ + _____ = _____ M&M's
blue green

_____ + _____ = _____ M&M's
orange red

Figure 5.

Name _____ personal response

The color M&M I like the most is _____ because

_____

_____.

_____

If I could choose a new color for an M&M, I would choose _____ because

_____

_____

_____.

Figure 4:

MSPAP Activity

Scenario

Your parents built a clubhouse for you and your friends. You love the clubhouse and play in it everyday! However, you and your friends decide that the clubhouse does need some things inside to make it better. Your parents won't buy anything but they do have some materials you can use. You need to decide with your friends what to make for inside your clubhouse. List your best 3 ideas on the chart and decide as a group on one that you can make. You will be able to use boxes, scissors, glue, staplers, construction paper, tubes, and yarn. You must work as a group to make one clubhouse item and use at least 2 boxes. Work cooperatively and have fun!

Figure 5:

Name _____

Date _____

Writing Prompt: Writing To Inform

Pretend that you have been hired to work as a travel agent for the summer. As part of your job, you have been asked to design a brochure about Norway for American tourists. Before completing this task, you will need to read an article concerning Norway. When reading, you may want to think about the people, places and interesting facts pertaining to Norway.

Remember your brochure will be read by Americans who are interested in traveling to this beautiful country. Therefore, you must be sure your writing is clear and complete and that you have used correct capitalization, word usage, punctuation, and spelling.

Now you will Read To Be Informed. When reading to be informed, you must do the following:

- Think about what you want to learn or find out from the material.
- Skim to find out how the author has chosen to present the information.
- Look for aids the author has provided: tables, illustrations, diagrams, boldface print, underlining captions or glossaries.
- Pay attention to titles, and subheadings or subtitles.
- Pause during your reading to organize the information.

When Writing to Inform, you must do the following:

- Think about what the person you are writing to needs to learn about the topic or subject.
- Put information in a logical order.
- Use examples, definitions, and descriptions to make the information clear to the reader.

Figure 6:

Name:_____ Date:_____

Prompt

Write a letter to Mr. Morrow. Persuade Mr. Morrow to add a new piece of playground equipment to the playground. When you write your letter, think about the piece of playground equipment that you would like, how it is like or different from the other equipment on the playground, who would like to play on it, and how he could raise money to purchase the equipment.

Topic _____

Audience _____

Purpose _____

Figure 7:

Name:_____

Date:_____

Scoring Rubric: Summary

Score Point 4

*Completely addresses all parts of the Story Map
*Events are discussed in a logical order
*Details are given which briefly describe each event
*Paragraph is well developed (contains a topic sentence, concluding sentence and uses transition words correctly)
*Contains consistently correct CUPS (Capitals, Usage, Punctuation, Spelling)

Score Point 3

*Completely addresses all parts of the Story Map
*Events are discussed in a logical order
*Paragraph is developed (contains a topic sentence and concluding sentence but only some transition words are used correctly)
*Contains generally correct CUPS (has some errors)

Score Point 2

*Partially addresses the Story Map
*Events are not discussed in a logical order
*Paragraph is partially developed (contains a topic sentence or a concluding sentence and transition words are not used correctly)
*Contains noticeable errors in CUPS

Score Point 1

*Minimally addresses the Story Map
*Events are not discussed in a logical order
*Paragraph is not developed (contains neither a topic sentence or a concluding sentence and transition words are not used)
*Mostly contains errors in CUPS

Score Point 0

*Blank: No response
*Response does not address the topic
*Unscorable: Response cannot be read

Figure 8:

RUBRIC SCORING

3= ALL 5 PARTS OF UMBRELLA COMPLETED WITH A CORRECT WRITTEN INTERPRETATION OF EACH PART.

2= 3 OR MORE PARTS OF UMBRELLA COMPLETED WITH A CORRECT WRITTEN INTERPRETATION OF THREE PARTS.

1= 1 OR MORE PARTS OF UMBRELLA COMPLETED WITH A CORRECT WRITTEN INTERPRETATION OF TWO PARTS.

0= DID NOT ATTEMPT ASSIGNMENT OR ALL INFORMATION IS INCORRECT.

Figure 9.

GRADING CRITERIA

This book report project will be graded in three parts. These parts include your written portion of the project, your presentation, and your puppet.

WRITTEN CRITERIA

3 All written tasks are neatly completed.
Correct punctuation and capitalization are used.
Award design completed and colored.

2 Two written tasks are neatly completed.
Correct punctuation and capitalization are used.
Award design completed and colored.

1 One written task is neatly completed.
Some correct punctuation and capitalization used.
Award design completed.

ORAL PRESENTATION CRITERIA

3 Presenter speaks in a clear and loud voice.
Presenter looks at the audience.
Presenter can read what is written on their paper.

2 Presenter speaks in a clear and soft voice.
Presenter looks at the audience.
Presenter has some difficulty reading to the audience.

1 Presenter is difficult to understand.
Presenter looks at their paper.
Presenter does not know what is written on their paper.

PUPPET CRITERIA

3 Puppet clearly represents the Famous African American studied.
Puppet shows why that African American was famous.

2 Puppet represents Famous African American studied.
Puppet does not show why the African American was famous.

1 Puppet does not represent the Famous African American studied.

Figure 10:

RUBRIC FOR MSPAP ACTIVITY

3 POINTS

Correct problem and answer.
Correct explanation.
Paragraph with 5-7 sentences. Correct illustration with details and color.

2 POINTS

Correct problem and answer.

OR

Correct problem and explanation.

OR

Correct picture and explanation.
Paragraph had 3-5 sentences with color.

1 POINT

Correct problem.

OR

Correct explanation

OR

Correct paragraph with no illustration or color.